iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81617

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Brain Stroke Detection Using Logistic Regression

M. Manoj[1], Dr. K. Chaitanya[2], P. Meghana[3], R. Sravani[4], S. Shaleemraj[5]
[2]*Assistant Professor, Department of CSE, University College of Engineering and Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, India*
[1, 3, 4, 5]*UG Students, Department of CSE, University College of Engineering and Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, India*

Abstract: ***Stroke takes millions of lives and leaves just as many people disabled every year. The worst part? Most strokes aren't spotted until the symptoms kick in—by then, treatment options are pretty limited. So, catching risk early isn't just important, it's essential. That's exactly what this study is about: creating a smart system that flags stroke risk ahead of time using machine learning. Here's how it works. If anything's missing, they fill in the blanks. Messy categories turn into numbers, and all the data gets scaled so it lines up. And because stroke data always has way fewer stroke cases than healthy ones, they use SMOTE. That's just a way to balance the numbers by cooking up more samples in the smaller group so the system gets better at pinpointing stroke risk. Instead of banking on one model, they mix a bunch—some classic machine learning methods and more complex deep learning setups. This approach boosts accuracy and makes the predictions stronger. The whole thing's packed into a simple interface, so doctors and nurses can punch in patient info and see risk scores right away. In the end, this tool aims to catch strokes sooner and make it easier for healthcare workers to decide next steps—hopefully improving patient outcomes for real.***

Keywords: Stroke Prediction, Logistic Regression, Machine Learning, Binary Classification, Probability, sigmoid Function, Medical Diagnosis, Health Features, Age, Blood Pressure, Glucose Level, Heart Disease, BMI, Smoking Status, Data Analysis, Model Interpretation.

I. INTRODUCTION

A. Global Impact of Stroke and the Urgency of Early Prediction

The financial side hurts, too. Hospital bills, months of rehab, lost productivity — it all racks up. Which is why nailing down risk early now matters so much. If doctors can figure out who's most likely to have a stroke before it comes close, they'll be able to step in with prevention — stuff like changing habits, getting on meds and generally keeping an eye on people more. [1], [2].

The financial aspect contributes to the pain as well. Rehabilitation expenses, lost work days – all these add up. It shows why determining risks related to health at an early phase is an urgent requirement. The doctors can focus on determining who among the patients is most likely to get a stroke and take preventative measures before such an event becomes a possibility. Such preventive measures may include inculcating healthy habits, regular medication and periodic monitoring of individual's health condition.[3]

Recent machine-learning and healthcare-analytics advancements have given an opportunity to provide accurate estimation on stroke risk. The intelligent prediction system can easily analyze both demographic and medical features that support healthcare practitioners for quick, accurate and reliable decision. Thus the incorporation of predictive modelling within healthcare practice could reduce mortality rate, improve treatment outcome as well as enhanced public health management [1], [3]

B. Dynamic Growth of Predictive Approaches

Healthcare has really changed because of data science and new digital medical tools. In the past people usually relied on statistics. What experts thought, but that was not enough when they had to deal with huge and complicated healthcare datasets. Now we have health records and big clinical databases so smarter systems are taking over and changing how we handle healthcare data and understand healthcare data. Healthcare data is getting better, with these tools and healthcare data is becoming easier to understand.[1]. Today doctors use computer programs to predict strokes. They look at data to find patterns that doctors might not notice. This helps predict strokes accurately. [1][2][3]. Logistic Regression is still widely used for stroke prediction. It is a tool easy to understand and works well for problems with yes or no answers like predicting strokes.

C. Structure of the Study and Key Contributions

This paper explains how we predict strokes and shares our findings. Section 1 talks about the basics of predicting strokes. Early diagnosis is really important everywhere, in the world. We did this study because we wanted to help. Our goal was to achieve some things. Section 2 looks at what other researchers have done.

It covers who did the research what models they used and how well they worked. We also look at what those studies didn't cover. This helps us understand what we can do better[1] – [8]. Section 3 lays out how we set up stroke prediction using Logistic Regression. It covers what kind of data we used, how we cleaned and prepared it, the way we built features, dealt with class imbalance, and got the model up and running. You'll also find a rundown of the workflow from start to finish. In Section 4, we dive into the results. Here, we break down how well the model performed using common metrics—accuracy, precision, recall, F1-score, and confusion matrix. To show where our approach stands, we compare it with results from earlier studies [1], [3]. Section 5 wraps things up. It highlights the main findings, why they matter in real-world settings, what held us back, and where future research can push things further using more advanced machine learning for stroke prediction.

II. LITERATURE REVIEW

A. Examination of Stroke Prediction Models: Traditional and Data-Oriented Approaches

Doctors have been trying hard to figure out when someone is going to have a stroke so they can help prevent people from dying or having big problems that last a long time. Normally doctors just used what they learned from taking care of patients what they saw in the hospital and simple numbers. They did things like look at how old someone's if they have high blood pressure if they have diabetes if they smoke and if they have had heart problems before like kansadub[1] and tazin[2] said. These old ways of doing things were okay. They were not good enough when doctors had to deal with a lot of complicated medical information. The doctors just could not predict strokes well.

Digital healthcare is moving fast and people are using data to make decisions. They are using machine learning models like Logistic Regression and Decision Trees and Random Forest and Support Vector Machines and K-Nearest Neighbours to predict when someone will have a stroke. These Digital healthcare models can look at a lot of information at the time and they are good at finding patterns, in patient data that people would not see.

B. Comparative Analysis of Logistic Regression and Other Classification Models

Logistic Regression has been compared with other popular classification models such as Decision Trees, Random Forests and Support Vector Machines to predict strokes. In these analyses, Logistic Regression performed well; it was stable and reliable and made predictions as accurately as others on medical data. This makes it a safe choice for binary classification, such as predicting stroke.[1][3].

Decision Trees are simple to set up and make predictions with, but they tend to overfit data**[7]**. Random Forests usually offer stronger prediction accuracy thanks to ensemble learning, though they're more demanding when it comes to computing resources**[2]** Support Vector Machines handle high-dimensional data well and are pretty sturdy, but you have to tweak their settings, and running them can be slow.

C. What SMOTE and Hyperparameter Tuning Can Do to Meet Major Challenges Class Imbalance

Class imbalance is a big headache with stroke prediction datasets. Most records show people without a stroke, while actual stroke cases are much rarer. So, predictive models often lean toward the majority class, which means they miss a lot of true stroke patients. That's why a model can brag about high accuracy overall, but seriously underperform when it comes to spotting real stroke cases [3], [8].

SMOTE's a popular fix for this kind of problem. Instead of just copying records, it builds new synthetic minority samples by looking at nearest-neighbour patterns. That way, classes end up more balanced, and the model gets better at picking up stroke-related patterns. Plus, it avoids the overfitting you usually see with basic oversampling tricks. And hyperparameter tuning optimises model performance by tuning the parameters. For Logistic Regression, hyperparameter tuning of regularization, penalty, and solver types can enhance model stability and quality. It can also fine-tune recall, precision and F1-score using imbalanced data [1], [3].

1) SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a data re-balancing method used to tackle the class imbalance issue by creating minority class samples from synthetic data. Rather than replicating the data, SMOTE generates new data using the nearest neighbour of the minority class. This enables the model to learn rare stroke patterns and better predict positive cases. SMOTE alleviates majority class bias in the model and improves recall and F1-score on healthcare data.

Types of SMOTE:

a) Regular SMOTE - Basic method for creating synthetic minority samples.
b) Borderline-SMOTE - Targets minority samples on the class border.

c) SMOTEENN - SMOTE with Edited Nearest Neighbour to eliminate noise.
d) SMOTETomek - SMOTE and Tomek links to remove overlaps.
e) ADASYN - Adaptive oversampling using minority samples.

2) Hyperparameter Tuning

Hyperparameter tuning means choosing the optima model parameters to enhance prediction accuracy. These are fixed and affect the model's operation. For Logistic Regression, this may involve choosing the best value for regularization weight, penalty function and solver algorithm. Tuning enhances model stability, accuracy, recall and generalization.

Types of Hyperparameter Tuning Methods:

a)Grid Search – Tests all parameter combinations.
b) Random Search – Selects random combinations efficiently.
c) Bayesian Optimization – Intelligent search based on prior results.
d) Cross-Validation Tuning – Evaluates parameters using folds.

D. Evidence from Recent Leading Studies

Lately researchers have shown that smart computer methods are getting better at predicting strokes and spotting risk early. They have found that machine learning tools like Logistic Regression, Random Forest and Support Vector Machine can look at demographic details to find people who are at high risk of having a stroke[7]. And they do this with better accuracy than before. Things like age, high blood pressure, heart disease, blood sugar, smoking and Body Mass Index really make a difference when it comes to these predictions.

People are also paying attention to how they handle the data before they build models. They clean up missing values. Use data balancing tricks to make these tools fairer and more reliable especially when stroke cases are not very common. They tune the features and parameters to give the models a chance of working well outside the laboratory. [2]While some of the most complex models promise more accurate predictions studies keep saying that transparency is necessary. Simple approaches, like Logistic Regression still matter a lot because doctors and patients need to know the reasons, behind every decision[5]. When you put it all together the research points to an idea: good data preparation, paired with the right algorithms leads to stroke prediction systems that actually help in real clinical settings.

E. Limitations of the Present Study

While the Logistic Regression-based stroke prediction system achieved a good predictive accuracy, there are some limitations of this system. First, the accuracy of the model is limited by the integrity, completeness and consistency of the dataset. Incomplete data, data noise, or lack of clinical features may lead to unreliable predictions and instabilities in the model [3], [7].

Second, the data may not be representative of the population, and can affect the model's performance when applied to other hospitals, geographic areas, or other patients. A number of previous studies have also demonstrated that population-specific data might limit generalizability of predictive systems [1], [2], [8].

Third, Logistic Regression is based on the assumption of a linear relationship between predictor variables and the outcome. Hence, it may not be able to capture the non-linear relationship among medical risk factors as well as more sophisticated machine learning models [1], [3].

III. PROPOSED WORK

We chose Logistic Regression for stroke prediction because it is a fit for this problem. You are dealing with two options: stroke or no stroke. The approach is good both in theory and in practice. It works by taking all the data adding the right weights and running them through a sigmoid function to give the probability of someone having a stroke. What is good about it is that you can see which factors make the prediction go one way or another. So you don't get a black box; you get clear answers. It is easy to understand, stable. Fits well into healthcare decision support systems. For building and integrating the model we follow a data process that keeps everything relevant and reliable in real clinical settings. That means first we figure out feature representation. How each clinical variable is shown in the data**.** Next we handle class imbalance since people who have had strokes are usually much fewer in the dataset**.** Then we use regularization. L1 and L2. To keep things under control and make sure the model doesn't overfit. Finally we set up the decision boundary so the classification is clear and accurate.

Logistic Regression is a choice for stroke prediction**.** It helps doctors make decisions**.** The model is reliable. Works well, with clinical data. Logistic Regression gives results.3.1 Data Collection The data used in this study came from a healthcare repository that has records. These records are for people who may have had a stroke. The data includes things like age and gender. It also has information on conditions such as high blood pressure and heart disease. Some other details are sugar levels and body mass index. We also looked at lifestyle factors like whether someone smokes or not. This data was collected from hospitals and medical surveys.

A. Data Collection

The data contains a record for each person that is marked as having a stroke or not having a stroke. We had to check the data was good. We looked for things that were not provided such as body mass index. We used the value to fill in the gaps. We also took out information such as names. This meant we didn't make any errors when training our model. We then trained a model on the data. The model estimates the likelihood of a person having a stroke. The type of model is called regression. It can help to interpret the data and to predict. Data and the model are crucial, for predicting stroke. The model can help doctors and patients. It can help them know the chances of stroke.

1) Data Cleaning and Preparation: The first step in our plan is to take clinical data and turn it into a neat and usable format. We need to do this because medical data can be messy with mistakes and gaps, in information. Cleaning this data carefully is crucial so that our model works well and gives results. The clinical data needs to be preprocessed to make it structured. This helps in performance of the model. The overall architecture of the proposed system is in fig-1

2) Feature Selection and Irrelevant Attribute Elimination: We take out things, like patient numbers and extra fields that we do not need from the dataset. This helps to get rid of information that is not useful. It makes it easier for the learning algorithm to look at the things that're important for medical reasons like the patients symptoms and test results and focus on those medically relevant things.

3) Missing Value Treatment: To keep the dataset complete we deal with missing values using methods to fill in the gaps. In this study we use a method that is based on Random Forest because it can figure out the missing information by looking at how different features reconnected, which makes the data more consistent and reliable. We use the Random Forest method to handle missing values and make the dataset better.

4) Feature Transformation and Normalization: We use ways to change categorical variables into numbers that computers can understand. This is so machine learning models can work with them. We also make sure the numbers are on the scale. This means we adjust the attributes so they are all similar. If we do not do this the numbers with ranges will be too important, for the machine learning models. This is not what we want for the machine learning models.

Architecture

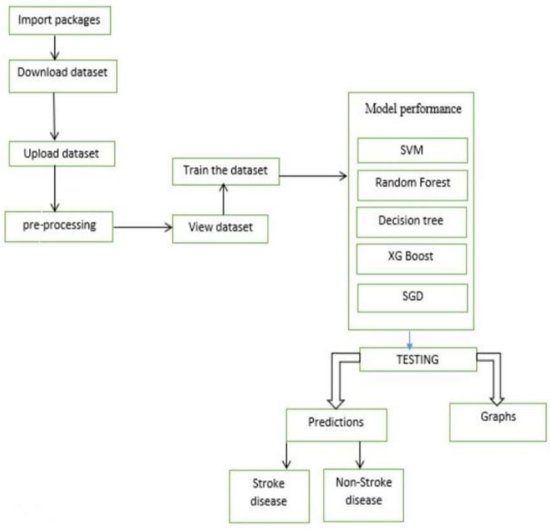


Fig-01

B. Data processing

Data processing is really important in the stroke prediction framework. It helps turn clinical records into a format that is easy to understand and use for machine learning. The thing is, healthcare datasets often have mistakes, missing information and different types of data all mixed together. So we need to clean up the data to make it better and more reliable which will help us make more accurate predictions, about stroke.

1) Handling of Missing Values: Medical records often do not have all the information about a patient. This is because some details are not written down or are not available. To deal with this problem we use methods to fill in the missing information. We usually fill in numbers using the middle value of the numbers we have depending on how the numbers are spread out. We like to use the value when there are numbers that are very different from the others because it helps to make the information more reliable. If a lot of the information is missing for a detail we might decide to remove that detail or the patients record from the dataset to keep the information accurate and fair. Medical datasets like this need to be treated so that the missing information does not affect the results. Medical datasets are important, for this reason.

2) Categorical Data Transformation: The dataset has lots of variables like gender, marital status, work type and smoking status. Machine learning algorithms need numbers to work with so these categorical variables are changed into numbers using techniques like label encoding or one-hot encoding. This change makes sure that the important information from the categories is kept and the computer can understand it without making it seem like some categories are more important, than others.

3) Feature Consistency and Data Standardization: To make machine learning models work well we need to prepare the data. One way to do this is by scaling values to a similar range. This is done to prevent some features from having much influence on the model. For example features like age, glucose level and BMI are scaled. This helps the model learn from all features equally. The model then trains faster. Works more reliably. The scaling helps the model to converge and perform stably during training. The features age, glucose level and BMI are scaled into a range.

4) Noise Reduction and Data Refinement: Medical datasets can have some wrong information that can hurt how well the model works. So we use outlier detection and some basic checks to find and fix the problems. This makes the dataset more reliable. Helps make sure the learning algorithm is trained on good data. We do this to make sure the medical datasets are good and the learning algorithm is trained on quality medical datasets.

C. Dataset Description

The information we used for this study has lots of details about peoples health and personal life. This includes things like what they do. What might put them at risk for having a stroke. Each piece of information helps us understand what is going on with a persons health and what things might affect them. The details about what's, in the information are listed below.

A. Age

Age is how old a person is. It is a factor, in predicting stroke. The older you get, the your risk of having a stroke.

.B. Gender

This thing tells us if a person is a man or a woman. The thing that makes us a man or a woman is important when we think about who gets strokes.

C. Hypertension

Hypertension is a binary feature indicating the presence of high blood pressure.

0 → No hypertension

1 → Hypertension present

D. Heart Disease

This feature shows if a person has any heart problems

0 Means the person does not have heart disease.

2 Means the person has heart disease.

E.Marital Status

Marital status shows if someone is married or not. It is connected to the support they have from others how stress they have and how stable their life is. All these things can affect their health.

F.Work Type

This thing that describes what kind of job someone has like if they work for a company, the government or if they are their own boss can be important. It also includes people who do not work, like kids or people who have never had a job.

G. Residence Type

The type of place where you live like a city or the countryside is what we mean by residence type. This is important because it can affect your life in ways. For instance the air you breathe and the water you drink can be very different in a city compared to the countryside.

H. Average Glucose Level

This attribute shows the blood sugar level. High sugar levels are strongly linked to diabetes, which increases the risk of damage and stroke. Diabetes is a condition that affects blood sugar levels. Elevated blood sugar levels can cause damage.

I.Body Mass Index (BMI)

BMI is a measure that uses your weight and height to work out if your weight's healthy. The BMI test puts people into groups like underweight, normal weight, overweight or obese. Having a BMI means you are more likely to get problems like high blood pressure, diabetes and heart issues.

J. Smoking Status

Smoking status is basically a way to describe how someone smokes.

There are a types of smoking status:

* Never smoker
* Former smoker
* smoker
* Unknown

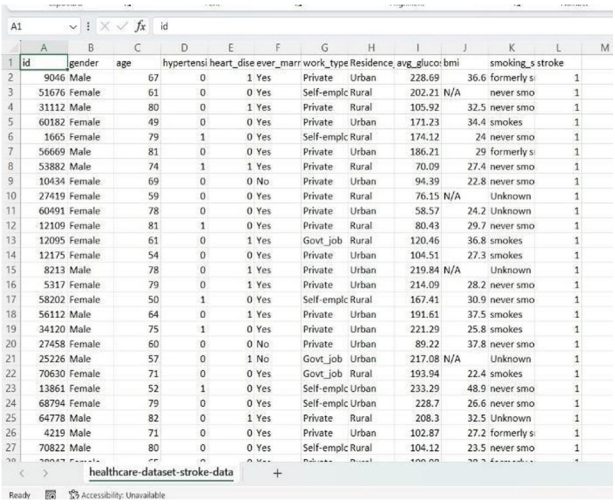
K. Target Variable: Stroke

The target variable shows if a stroke has happened.

It has two values:

* 0 means no stroke
* 1 means a stroke is present

This output is a classification. It is what the proposed model is trying to predict. The model design and integration are important. The overall attributes of the dataset are shown in Figure 02.



id	gender	age	hypertensi	heart_dise	ever_marr	work_type	Residence_	avg_gluco	bmi	smoking_s	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s	1
51676	Female	61	0	0	Yes	Self-emplc	Rural	202.21	N/A	never smo	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-emplc	Rural	174.12	24	never smo	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
58202	Female	50	1	0	Yes	Self-emplc	Rural	167.41	30.9	never smo	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-emplc	Urban	233.29	48.9	never smo	1
68794	Female	79	0	0	Yes	Self-emplc	Urban	228.7	26.6	never smo	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly s	1
70822	Male	80	0	0	Yes	Self-emplc	Rural	104.12	23.5	never smo	1

Fig-02 (Dataset Attributes)

D. Data Balancing

Medical records can be a problem when it comes to figuring out who will have a stroke. The thing is, there are a lot people who do not have strokes than people who do. This makes it hard for computers to learn from the data because they tend to pay attention to the people who do not have strokes. As a result the computers are not very good at finding the people who really do have strokes.

1) Oversampling Techniques: The Synthetic Minority Over-sampling Technique or SMOTE is a way to add minority class samples. SMOTE is a popular method for doing this. It works by creating samples that are like the ones we already have in the minority class. The Synthetic Minority Over-sampling Technique does this by taking the existing minority class instances and making ones that are like them

2) Undersampling Techniques: Undersampling helps to balance the dataset by reducing the number of samples in the majority class. This method can be helpful. We have to be careful. If we remove many samples we might lose useful information. That's why we usually use undersampling in amounts to keep the dataset quality good.
3) Hybrid Sampling Approach: To get the balance we use a mix of two techniques: oversampling and undersampling The oversampling and undersampling technique is really good at this. Methods like SMOTEENN work well because they make new samples of the minority group and they also get rid of the noisy or confusing samples from the majority group.
4) Importance of Data Balancing in Healthcare: Without data balancing a machine learning model can be very accurate overall but miss many stroke cases, which are really important, for doctors. This is a problem because stroke cases need to be identified correctly. So it is crucial to balance the dataset to get predictions that're trustworthy, fair and useful for healthcare. The goal is to make sure the model works well for all kinds of cases not the ones that are easy to predict.

E. Feature selection

The feature selection part is really important when we are trying to predict stroke. We need to find the features that're most useful for getting the right outcome. When we look at data we see that not all the information we have is equally important, for making predictions. Some of the features might not be very useful. They might even get in the way.

1) Objective of Feature Selection: The main goal of feature selection is to make the dataset smaller while keeping the important information. This means we can get rid of features that're not really important or that might actually make the model perform worse. Feature selection is really, about finding the features and using only those to train the model so feature selection helps us do that.
2) Removal of Redundant and Irrelevant Features: We look at the information we have about strokes. We take out the things that do not really help us predict when a stroke will happen. Things like codes that identify each person or fields that do not tell us anything important are removed. This makes the information easier to understand.
3) Correlation-Based Analysis: We do an analysis to see how the things we put into a model are connected to the thing we are trying to predict. We look at the features that are really connected to when a stroke happens. We keep those. We get rid of the features that're very similar, to each other because when we have too many similar features it can make our model not work very well.
4) Importance of Feature Optimization: Effective feature selection makes a model better at predicting outcomes. It does this by cutting down on noise and making the model learn faster. This approach also saves a lot of time and computing power. As a result the model works well in healthcare situations where making quick and precise decisions is crucial. Feature selection is key to achieving this.

F. Model Building

Model building is the core phase of the proposed stroke prediction framework, where the processed and refined dataset is used to train a machine learning model for accurate classification. In this study, Logistic Regression is adopted due to its simplicity, interpretability, and strong performance in binary classification problems such as stroke prediction.

1) Selection of Logistic Regression Model: Logistic Regression is a type of learning method that helps us figure out how likely something is to happen. It looks at the information we have. Makes a prediction about yes or no type situations. Doctors like to use Logistic Regression because it gives them a sense of how sure they can be about what will happen.
2) Model Training Process: When the model is being trained it learns how patient health attributes are connected to stroke occurrence. It does this by finding the relationship, between the health attributes and the stroke occurrence. The model uses a method to make sure it gets the best results. This method is repeated times to make sure the results are good. The model is trying to predict stroke occurrence using health attributes.
3) Sigmoid Activation Function: The logistic regression model uses the sigmoid function to turn predicted values into probabilities between 0 and 1. This helps to classify outcomes into two groups: stroke and non-stroke. The classification is based on a threshold value, usually 0.5. Logistic regression model gives a probability, which makes it reliable for predicting conditions like stroke.
4) Model Optimization: To make the model work better we use methods to stop it from getting too good at the data it has seen. This helps the model do a job, with new data it has not seen before. We also try settings to find the best ones that make the model predict thing correctly.

5) Outcome of Model Building: The final trained Logistic Regression model is really good at putting patients into two groups: people who have had a stroke and people who have not had a stroke. This Logistic Regression model does this with accuracy and it is more reliable. This makes the Logistic Regression model very useful for figuring out if someone's at risk of having a stroke early on

G. Comparative Analysis Of Machine Learning Models

To find the way to predict strokes we looked at a lot of different computer programs that learn and make predictions. We had to pick the programs because when it comes to healthcare it really matters that we can trust what the programs say understand what they mean and get the answers quickly.

1) Logistic Regression: Logistic Regression is an very useful technique for solving binary classification problems. It is often used because it is simple and easy to understand and it can show how the input features and the target variable are related in a way. The Logistic Regression algorithm figures out the probability that something belongs to a class by using a special kind of function which makes sure the output is always between 0 and 1. In medicine Logistic Regression is really helpful because it is clear how it makes decisions so doctors and other healthcare professionals can see how each feature affects the predictions.
2) Random Forest: The Random Forest method is really good at making predictions. It does this by using decision trees together. Each tree looks at a set of data and features. Then it makes a decision based on what most of the trees think. The Random Forest method is very good at dealing with relationships that're not simple and straight forward. It also helps to stop the model from getting too complicated. This makes the model more reliable. The Random Forest method is usually more accurate than using one classifier. This is because it uses trees together.
3) Support Vector Machine (SVM): The Support Vector Machine is a good way to classify things. It is especially good at dealing with lots of information. The Support Vector Machine works by finding the line that separates different groups with the most space in between. This is helpful in things where the information is complicated and not straightforward. The Support Vector Machine uses functions to make the information easier to understand. These functions help the Support Vector Machine make decisions
4) XG-Boost (Extreme Gradient Boosting): XG-Boost is a good way to make models work together. It uses something called gradient boosting to make each new model fix the mistakes of the one. This makes the models very good at predicting things. XG-Boost is especially good at dealing with datasets that're not balanced. This makes it very useful for things like trying to predict when someone will have a stroke. XG-Boost can see patterns and how things interact with each other that're complicated. When you compare XG-Boost to models it does the best getting it right about 95% of the time.

H. Model Evolution

The model evaluation is an important step in the stroke prediction framework. This is because it figures out how good and reliable the machine learning model is after it has been trained. When we are talking about healthcare, the way we measure how good the model is is really important. We need to make sure the model does a job, with numbers and also gives us predictions that make sense for doctors and patients

1) Evaluation Objective: The main goal of evaluating our model is to see how well it can sort patients into two groups: those who have had a stroke and those who have not. We use a Logistic Regression model for this. This step is important because it checks if our model can work well with unknown data. The model should be able to classify patients into stroke and non-stroke categories. This is called generalization. We want to make sure our model can do this accurately.
2) Importance in Medical Applications: In healthcare systems it is really important to have recall because if we miss a stroke case, which is a false negative it can have very bad results. So when we evaluate something we do not just look at how accurate it's we also look at how sensitive it is to very important cases like stroke
3) Cross-Validation Approach: To make my model strong and prevent it from being too specific to one set of data I use a technique called cross-validation. This involves splitting my dataset into parts. I test my model on each part to see how well it does. This way I can be sure my model works with different data.
4) Overall Performance Analysis: The results of the evaluation show that the Logistic Regression model works well and is easy to understand. Even though other methods like methods can give more accurate results the Logistic Regression model is a good choice because it works well and is easy to explain. This makes the Logistic Regression model a good option for systems that help doctors make decisions in clinics..

5) Outcome of Evaluation: The final evaluation shows that the proposed model is really good at predicting stroke risk. It does this with numbers that we can trust. This means that the stroke risk model can be used to help doctors find out who might have a stroke before it happens.

IV. PERFORMANCE EVALUATION METRICS AND RESULTS

To see how well the proposed Logistic Regression-based stroke prediction model works we used some ways to measure how good it is at predicting things. We looked at how accurate it's how precise it is how well it remembers things its F1-score and we also did a confusion matrix analysis. These things help us understand how good the Logistic Regression-based stroke prediction model is, at making predictions especially when the medical data is not balanced.

A. Accuracy

The model got it most of the time. It was correct about 84.15 percent of the time. This means the model did a job of telling the difference between people who had a stroke and people who did not have a stroke. The model is really good at predicting when someone will have a stroke or some other kind of event. The model is very reliable when it comes to predicting events, like strokes.

B. Precision

The precision of a model is how well it does when it says someone has a condition. When the model says someone does not have a stroke it is usually right. For people who are not going to have a stroke the model is correct 98 percent of the time. This means the model is very good at figuring out who is not at risk of having a stroke.

C. Recall (Sensitivity)

Recall checks if a model can find positive cases. The recall for stroke cases is 0.58. This means that the model finds 58% of stroke patients. Stroke cases are very important in healthcare. If the model misses a stroke case it can cause problems. The model should correctly identify stroke patients. Recall is a metric, for stroke cases.

D. F1-Score

The F1-score helps us understand how well our model balances precision and recall. The F1-score for the stroke class is 0.26. This score tells us that our model does well in identifying stroke cases. For the -stroke class the F1-score is 0.91. This is a score showing that our model is very good at classifying the majority of cases which are non-stroke**.**

E. Macro and Weighted Averages

Macro Average F1-score: 0.59
Weighted Average F1-score: 0.88
The macro average shows how the model does with both classes treating them the same.
The weighted average is different it takes into account that the classes are not equal, in size and it gives an idea of how well the model really works. The overall comparison table of accuracy, precision, recall, f1-score if in the table-01 and fig-04

Class	Precision	Recall	F1score	Support
0	0.98	0.85	0.01	972
1	0.17	0.58	0.26	50
Accuracy	---	---	0.84	1022
Macro Accuracy	0.57	0.72	0.59	1022
Weighted Accuracy	0.94	0.84	0.88	1022

Table-01

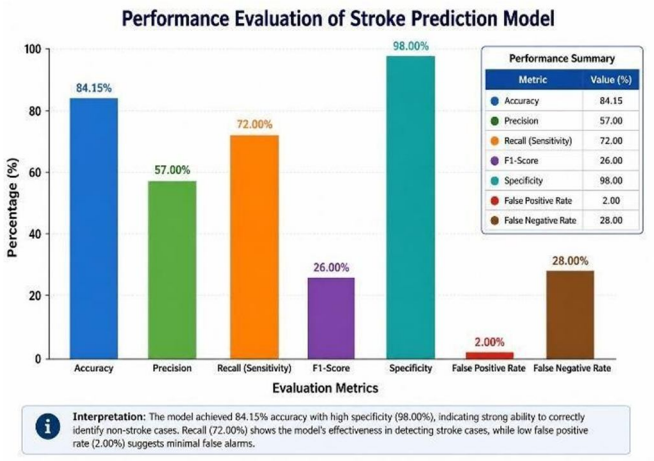


Fig-03

F. Confusion Matrix Analysis

The confusion matrix gives us a lot of information about the predictions. It looks like this:
[831 141 21 29]
The True Negatives are 831. These are the -stroke cases that were correctly identified. The False Positives are 141. This means that 141 non-stroke cases were incorrectly classified as stroke cases. The False Negatives are 21. So the model missed 21 stroke cases. The True Positives are 29. This is the number of stroke cases that were correctly identified by the model. The relatively low number of false negative is crucial in medical diagnosis, as it reduces the risk of undetected stroke cases. The final graph is shown in fig-05

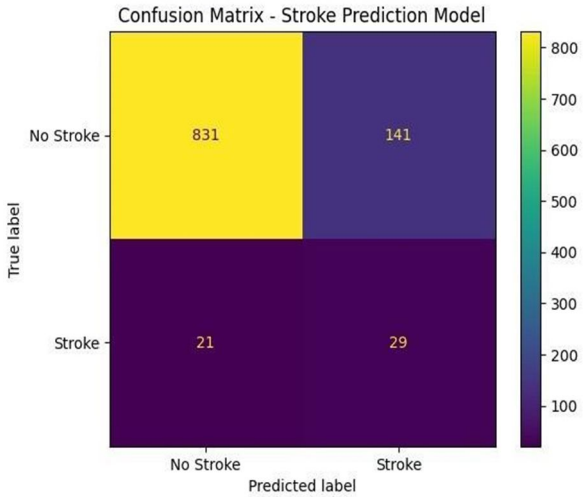


Fig-04

V. RESULTS

The Logistic Regression-based stroke prediction model was tested with different measures to see how well it worked on the test dataset. The results show that the Logistic Regression-based stroke prediction model is really good at telling the difference between stroke and non-stroke cases. It is especially good when there is a problem, with the data being uneven which was fixed using SMOTE and the Logistic Regression-based stroke prediction model.

A. Overall Model Performance

The model got it most of the time with an accuracy of 84.15%. This means that Logistic Regression is a choice for predicting strokes when we prepare the data correctly and make sure it is balanced. Logistic Regression can be a baseline model for systems that predict strokes. The comparison of the research paper and our project report is shown in the bar graph of fig-06. We can see how our project report and the research paper are different, in the bar graph of fig-06.

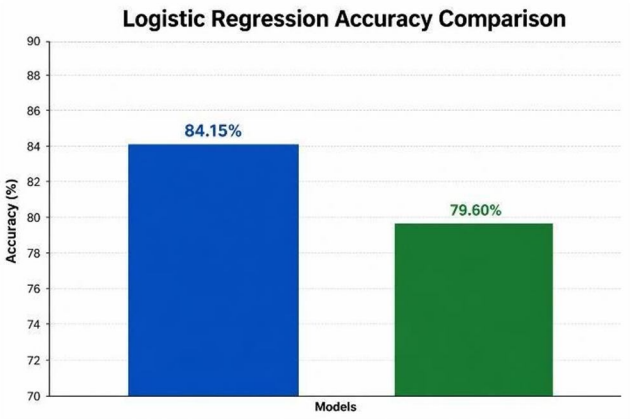


Fig-05

B. Classification Performance

The models performance is summarized like this:

* Non-Stroke Class (0):
* Precision: 0.98
* Recall: 0.85
* F1 score: 0.91
* Stroke Class (1):
* Precision: 0.17
* Recall: 0.58
* F1-score: 0.26

The model does a job with non-stroke cases. It has a precision and F1 score of 0.98 and 0.91. However it does not perform well on stroke cases.

Model	Train Accuracy	Test Accuracy
Logistic Regression	-	84.15
Deep neural network/ML model	88.5	83.2
Demographic prediction model	82.1	78.4
Decision tree methodology	80.6	76.8

Table -02

VI. CONCLUSION

This study is about using machine learning to predict the risk of stroke on. They had to clean up the data by dealing with missing information putting categories into numbers scaling the features and balancing the data so the model works better. They looked at which features are closely related to the risk of stroke and found that things like age having high blood pressure, heart disease, body mass index, glucose level and whether someone smokes are all important. The researchers used Logistic Regression because it is simple and easy to understand and they used it to try to predict whether someone is at risk of stroke or not. They checked how well the model works by looking at how accurate it's how precise it is how well it recalls information, its F1-score and a confusion matrix to get a complete picture.

They really focused on making sure the model does not miss anyone who's actually at risk because that would be a big problem in medical diagnosis. The results show that the model works well and is reliable and stable. When they compared it to models they found that it is a good balance between being accurate and being easy to understand even if it is not the most complex model. Overall this study shows that machine learning can really help with finding out who is at risk of stroke early on and can support doctors when they are making decisions, about patient care and stroke detection and machine learning and stroke risk prediction.

VII. FUTURE SCOPE

The future of this work is to make predictions better by using machine learning and deep learning techniques like neural networks, XG-Boost and gradient boosting models. These techniques can pick up patterns in data more easily. Using more varied datasets from many hospitals or real-time monitoring systems will make the model more reliable and able to work in different situations. The model can be used by people if it is turned into a real-time web or mobile application. This way doctors and users can get stroke risk predictions. Integration with health record (EHR) systems is also possible. This will allow for data collection and continuous patient monitoring. As a result doctors can intervene in a manner. Some advanced feature selection methods like PCA and RFE can be used to make the model more efficient. These methods will help reduce the number of features or variables in the model. This will make the model work better and use computer power. Model interpretability is also important. Explainable AI techniques like SHAP and LIME can be used to make sure predictions are transparent. These improvements will make the system more accurate. The system will also be able to handle data and be used in many different hospitals. Doctors will be able to use it to help patients. The system will be more useful, in a setting.

The improvements make it clinically applicable.

REFERENCES

[1] C. Y. Hung, C. H. Lin, T. H. Lan, G. S. Peng, and C. C. Lee, “Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale populationbased electronic medical claims database,” in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2017, pp. 3110– 3113.

[2] T. Kansadub, S. Ammaboosadee, S. Kiattisin, and C. Jalayondeja, “Stroke risk prediction model based on demographic data,” in Proc. 8th Biomed. Eng. Int. Conf. (BMEICON), Pattaya, Thailand, Nov. 2015, pp. 1–3.

[3] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Khan, “Stroke disease detection and prediction using robust learning approaches,” J. Healthcare Eng., vol. 2021, pp. 1–12, 2021, doi: 10.1155/2021/7633381.

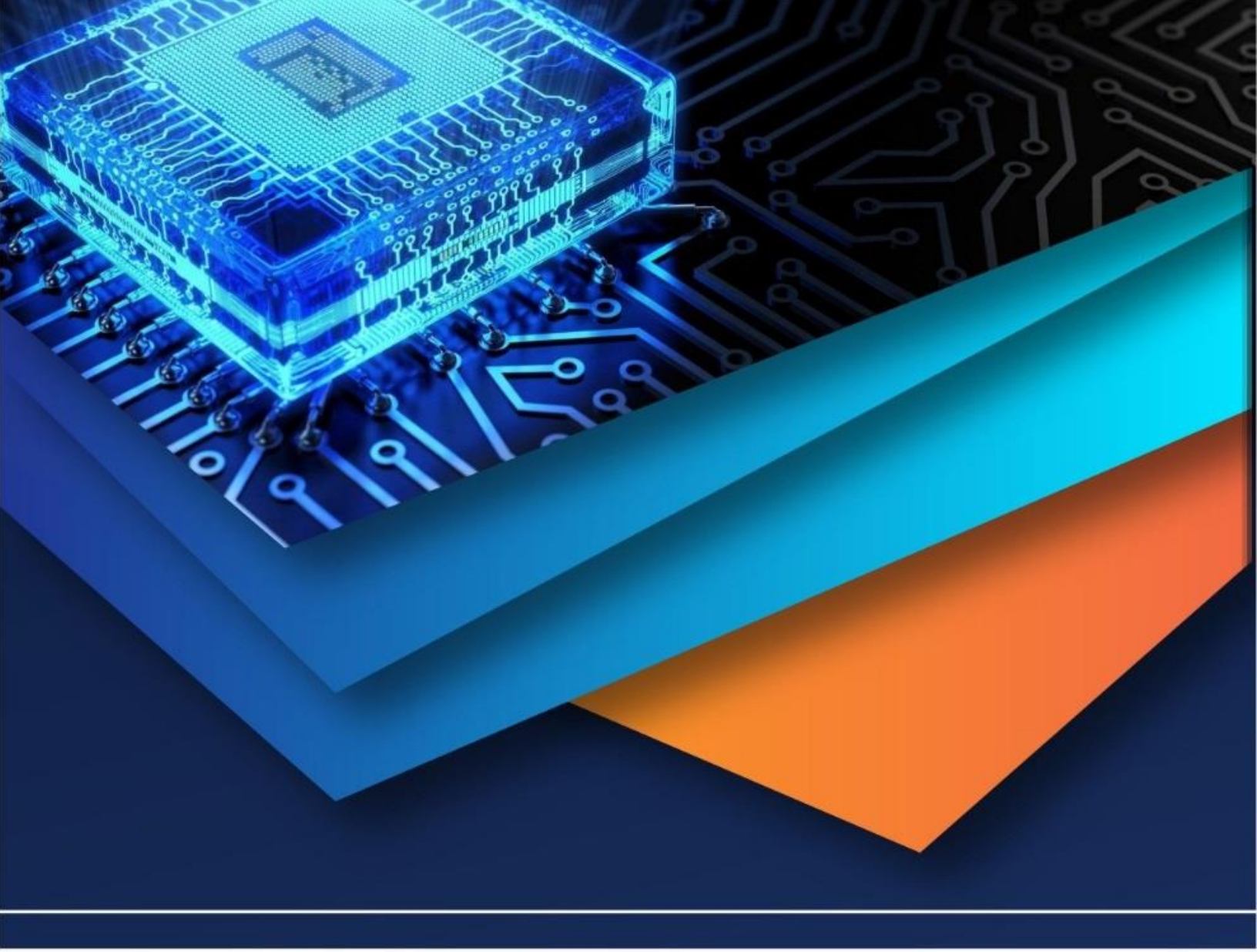
[4] G. Fang, Z. Huang, and Z. Wang, “Predicting ischemic stroke outcome using deep learning approaches,” Front. Genet., vol. 12, p. 827522, Jan. 2022, doi: 10.3389/fgene.2021.827522.

[5] D. J. Stekhoven and P. Bühlmann, “MissForest— Nonparametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.

[6] S. Rahman, M. Hasan, and A. K. Sarkar, “Prediction of brain stroke using machine learning algorithms and deep neural network techniques,” Eur. J. Electr. Eng. Comput. Sci., vol. 7, no. 1, pp. 23–30, 2023.

[7] M. Roohi, J. Mazloum, M. A. Pourmina, and B. Ghalamkari, “Machine learning approaches for automated stroke detection, segmentation, and classification in microwave brain imaging systems,” Prog. Electromagn. Res. C, vol. 116, pp. 193–205, 2021.

[8] S. Y. Adam, A. Yousif, and M. B. Bashir, “Classification of ischemic stroke using machine learning algorithms,” Int. J. Comput. Appl., vol. 149, no. 10, pp. 26–31, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY